A BRIEF INTRODUCTION TO DIFFUSION MODELS

Shuo Liu
Computer Science
Northeastern University
shuo.liu2@northeastern.edu

ABSTRACT

This article introduces mathematical foundations of diffusion models.

1 DIFFUSION MODELS

The diffusion model originates from the molecular motion of thermodynamics, which degrades the data distribution by gradually adding Gaussian noise and then uses a learnable model to recover it. Compared to other generative models like Generative Adversarial Networks (GANs) Goodfellow et al. (2014), Variational Autoencoders (VAEs), and flow models, the diffusion model is learned by one network with high-dimensional latent variables in flexible architecture, avoiding unstable training, mode collapse, and surrogated loss.

1.1 PROBLEM FORMULATION

Given a dataset $X = \{x_1, x_2, \cdots x_n\}$, where each sample is drawn independently from an underlying data distribution $p(x)$, the goal of generative learning is to fit a model to the data distribution such that we can synthesize new data points at will by sampling from the distribution.

1.2 NOISE CONDITIONAL SCORE-BASED NETWORK (NCSN)

The Score Matching with Langevin Dynamics (SMLD) is one of the first representative works of the diffusion model Song & Ermon (2019), which utilizes iterative *Langevin dynamics* to draw the samples. Langevin dynamics provides a Monte Carlo Markov Chain (MCMC) procedure to sample from a distribution $p(x)$ with only its score function $\nabla_x \log p(x)$ Welling & Teh (2011). It initializes the chain from an arbitrary prior distribution $x_0 \sim \pi(x)$ and iterates as follows,

$$x_{i+1} \leftarrow x_i + \epsilon \nabla_x \log p(x) + \sqrt{2\epsilon} z_i, \tag{1}$$

where $z_i \sim \mathcal{N}(0, I)$. When $\epsilon \to 0$ and $K \to \infty$, x_K obtained from Equ. 1 converges to a sample from $p(x)$ under some regularity conditions.

In the perturbing process, SMLD adds a sequence of multi-scale random Gaussian noises to the original data distribution $p(x)$,

$$p_{\sigma_k}(x) = \int p(y)\mathcal{N}(x; y, \sigma_k^2 I)dy, k = 1, 2, \cdots, K, \tag{2}$$

The SMLD then approximates the score function $\nabla_x p_{\sigma_k}(x)$ of each noise-perturbed distribution by training a Noise Conditional Score-Based Network (NCSN) $s_\theta(x, k)$. The score-matching objective of an NCSN is to minimize the weighted sum of Fisher divergences for all noise scales,

$$L(\theta) = \sum_{k=1}^{K} \lambda(k) \mathbb{E}_{p_{\sigma_k}(x)}[||\nabla_x \log p_{\sigma_k}(x) - s_\theta(x, k)||_2^2], \tag{3}$$

where $\lambda(i) \in \mathbb{R}_{>0}$ is a positive weighting function ($\lambda(i) = \sigma_k^2$). After obtaining the score-based model $s_\theta(x, k)$, we can produce the samples from it by running the Langevin dynamics for $k = K, K-1, \cdots, 1$ in sequence (so-called annealed Langevin dynamics).

1.3 DENOISING DIFFUSION PROBABILISTIC MODEL (DDPM)

Compared to NCSN which uses the score-based function (though convertible), the Denoising Diffusion Probabilistic Model (DDPM) reconstructs the samples from the noise according to *variational inference*.

In the forward chain, the DDPM gradually perturbs the raw data distribution $x_0 \sim q(x_0)$ to converge to the standard Gaussian distribution $q(x_k)$,

$$\begin{aligned} q(x_k|x_{k-1}) &= \mathcal{N}(x_k; \sqrt{1-\beta_k}x_{k-1}, \beta_k I), \\ q(x_{1:K}|x_0) &= \prod_{t=1}^{T} q(x_k|x_{k-1}), \end{aligned} \tag{4}$$

where $\beta_k \in (0, 1)$ is the coefficient of noise added at step t. On the other hand, the reverse chain seeks to train a parameterized Gaussian transition kernel with θ to recover the data distribution,

$$\begin{aligned} p_\theta(x_{k-1}|x_k) &= \mathcal{N}(x_{k-1}; \mu_\theta(x_k, k), \sigma_\theta(x_k, k)I), \\ p_\theta(x_{1:K}|x_k) &= \prod_{t=1}^{T} p_\theta(x_{k-1}|x_k). \end{aligned} \tag{5}$$

Our objective is to estimate the maximum likelihood of original distribution $p_\theta(x_{0:K})$ by maximizing the variational lower bound $\mathbb{E}_q[\frac{p_\theta(x_{0:K})}{q(x_{1:K|x_0})}]$. After parameterization Ho et al. (2020), we optimize the loss function,

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,I), k} ||\epsilon - \epsilon_\theta(x_k, k)||_2^2, \tag{6}$$

where ϵ_θ estimates the noise input. Once trained, we can sample x_0 from Equ. 5.

1.4 SCORE-BASED GENERATIVE MODEL (SGM)

The Score-based Generative Model (SGM) using *Stochastic Differential Equation* (SDE) Song et al. (2021) describes the diffusion process in continuous time steps with a standard Wiener process, which unifies NCSN and DDPM. The forward diffusion process in infinitesimal time can be formally represented as

$$dx = f(x, k)dk + \sigma(k)dw, \tag{7}$$

where w denotes a standard Winener process and $\sigma(\cdot)$ denotes the diffusion coefficient, which is supposed to be a scalar independent of x. The reserve SDE describes the diffusion process running backward in time to generate new samples from the known prior x_k Anderson (1982), which is

$$dx = [f(x, k) - \sigma(k)^2 \nabla_x \log p_k(x)]dk + \sigma(k)d\bar{w}, \tag{8}$$

which incorporate a backward induction score function $\log p_k(x)$. Similar to NCSN, the score function can be approximated using a step-dependent score-based model $s_\theta(x, k)$ with the score-matching optimization objective Equ. 3.

REFERENCES

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. ISBN 9781713829546.

Yang Song and Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.